UGent participation in the Microblog Track 2012

Thong Hoang Van Duc, Thomas Demeester, Joannes Deleu, Piet Demeester, Chris Develder
Dept. of Information Technology - IBCN
Ghent University - iMinds, Ghent, Belgium
{thoang, tdmeeste, jdeleu, piet.demeester, cdvelder}@intec.ugent.be

ABSTRACT

In this paper, we describe the search system, developed at Ghent University for the TREC 2012 Microblog Track in order to rank Twitter messages or 'tweets' from a fixed corpus in response to a number of search requests. Our system ranks the tweets based on a Logistic Regression classifier trained with data from the Microblog Track 2011. The features used for training the classifier include local tweets features, but also, query expansion and tweet expansion features, based on external Web data, which appear to significantly improve results.

Categories and Subject Descriptors

D.3.3 [**Microblog Storage and Retrieval**]: Microblog Search and Retrieval – *retrieval models, information filtering, search process*

General Terms

Microblog Retrieval models

Keywords

Twitter, query expansion, document expansion, classification

1. INTRODUCTION

Microblog services such as Twitter, Tumblr, Jaiku, etc. have become major types of social media on the web. These blogging applications allow users to broadcast short messages, individual images, status updates, or video links as well as general web links, to share information with friends, family and the general public. Recently, people have become more interested in the ubiquitous communication medium of the microblog, than in "long" forms of communications (e.g., traditional blogs). Despite the amount of research in the microblog area in the past few years [13] [14], search and online ranking on microblogs have not yet been addressed extensively. Therefore, the TREC Microblog Track was initiated in 2011 to replace the previous Blog Track.

For TREC 2012, we participated in the real-time ad-hoc task of the Microblog Track. A corpus of more than sixteen million tweets, Twitter microblog messages, was fetched using the Twitter API, and 60 search topics and relevance judgments were provided by the track organizers [1]. In the real-time adhoc task, the user issues a query at a temporal reference point and is looking for tweets that contain the most *relevant* and *recent* information to the query. Hence, the system should answer a query by providing a list of relevant tweets ordered from latest to earliest, up until the time the query was issued. This search task leads to a number of interesting issues, specific to the nature of a microblog service, which became apparent while we developed our system. First, each microblog post is very short, by definition. Tweets are limited in length to 140 characters but may contain hyperlinks to a specific topics or users. Abbreviations, phonetically shortened terms, drop vowels, etc. [2] are often found in tweets. This frequently leads to a vocabulary mismatch problem between the query and Twitter messages. Second, users write tweets in various formats. Some microblog posts are carefully written and clear to read, whereas others are quite difficult to read. However, the links, properties (hashtag, retweets, etc.) of a low quality post may still produce valuable information. Finally, microblogs are in multiple languages for people all around the world. However, this track only considers non-English messages by default as non-relevant. Therefore, we had to filter out non-English tweets first.

For our system, we applied various techniques to retrieve more relevant tweets. In particular, we explored query expansion and tweet expansion. We applied Logistic Regression to model the relevance scores of the retrieved tweets, based on features that were extracted from the tweets themselves and some external data, in order to improve the accuracy of our search system with respect to traditional IR methods.

The paper is organized as follows. In Section 2, we briefly describe our retrieval system. In Section 3, we show our experimental results, based on feedback received from the TREC organizers. Finally, we summarize our findings and possible future work.

2. RETRIEVAL SYSTEM

As this is the first time we participate in the Microblog Track, we focus on establishing a baseline system that can be easily extended for addressing related research questions and for further developing our retrieval system in the future. These are the main elements of our baseline system:

- A language detector, used to filter the English tweets in Twitter.
- A Lucene [3] index, used for basic tweet scoring with respect to queries.
- Document expansion and query expansion, used to help overcome the vocabulary mismatch problem.
- A linear classifier (logistic regression) [4], used to combine multiple features into an overall relevance score.

Figure 1 briefly outlines architecture of our search system's architecture. We submitted four officials runs that are different combinations of these basic elements. The following sections will be a brief description of each building block.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
UGent participation in the Microblog Track 2012	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Ghent University - iMinds,Dept. of Information Technology - IBCN,Ghent, Belgium,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT

In this paper, we describe the search system, developed at Ghent University for the TREC 2012 Microblog Track in order to rank Twitter messages or ?tweets? from a fixed corpus in response to a number of search requests. Our system ranks the tweets based on a Logistic Regression classifier trained with data from the Microblog Track 2011. The features used for training the classifier include local tweets features, but also, query expansion and tweet expansion features, based on external Web data, which appear to significantly improve results.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**5**	

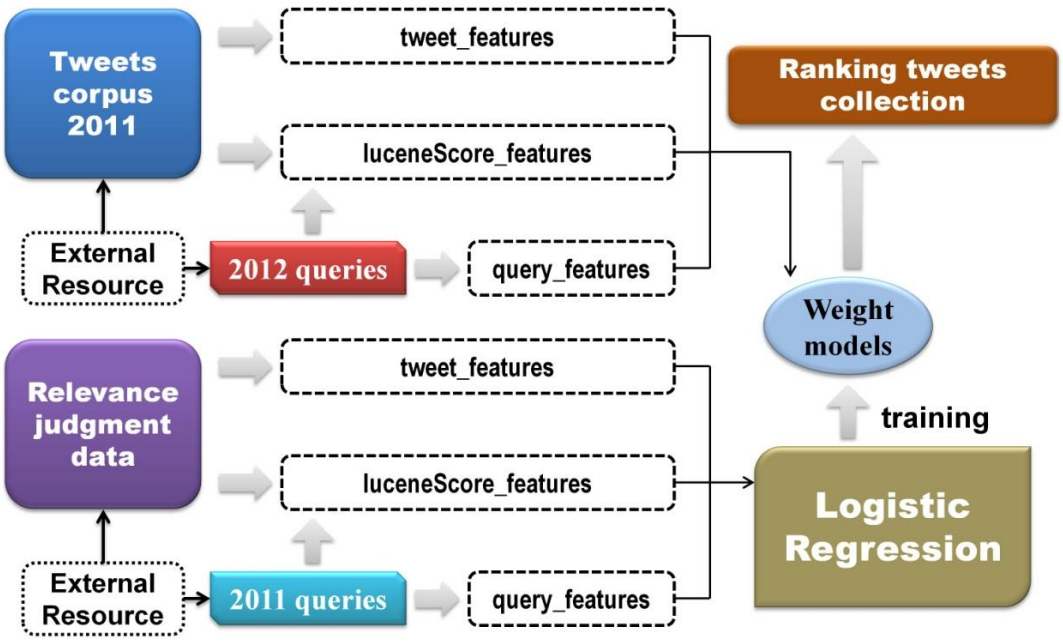

Figure 1. Overview of tweet ranking system

2.1 Preprocessing Data

According to the track guidelines, only English tweets are considered as potentially relevant. We therefore use the language detection tool LangDetect [5] to filter non-English tweets from the total tweets collection, as it is fast and accurate [12].

2.2 Text Scoring

One of the used features is the basic score that a standard IR system assigns to the tweets in response to the training queries from the 2011 track. The reason we chose for the Lucene search engine, is that it is a robust, powerful, free and flexible search toolkit. Note that we could just as well have used other search engines like Indri, Terrier, etc.

2.3 Document and Query Expansion

As indicated earlier, text-based messages in Twitter are limited to 140 characters. This leads to a vocabulary mismatch problem between queries and documents (tweets). We explore two common approaches to overcome the lexical gap between documents and queries: document expansion and query expansion.

Document expansion: By inspecting the judgments used for relevance evaluation in 2011, we find that most of the relevant tweets contain URLs in their post (around 94% of the highly relevant tweets and 80% of all relevant tweets, as opposed to 53% in the non-relevant tweets). This suggests that the presence of URLs can lead to valuable information for extending a large fraction of the tweets and queries. However, collecting the entire pages linked to by these URLs was judged unnecessary to build a baseline system. Not only would we need a general parser to detect possible relevant blocks of text in the result files, many of the URLs also pointed to graphics or multimedia data, which could not be directly used for our purposes. However, when we investigated the source files of these URLs, in most of the cases the page titles appeared to bring the essence of the page's content in well-chosen terms. Additionally, mostly informative (and therefore potentially relevant) tweets (e.g., press articles) contained links to pages with a clear title. For these reason, our crawler system just extracted the page title for the URLs. To avoid having to crawl data from each tweet that contains a URL in the whole collection, we chose the top 2000 tweets with the highest Lucene scores for each of the 2012 test queries. For this top 2000 (per query), the page titles related to the URLs were collected. After extracting the titles from the URLs, we just simply appended these to the original tweets to expand them. As expected, some of the titles could not be crawled because the website was blocked, spam or no longer existed. Table 1 shows the number of titles that were collected by our system for the training queries. *Ave.tweets* means the average number of tweets which are returned for each query, *ave.URLs* means the average fraction of URLs in the tweet collection, and *ave.title collected* are the fraction of tweets for which the page title was collected by our crawler system.

Table 1: Titles collection data

ave.tweets	ave.*URLs*	ave.titles collected
1379.72	0.47	0.24

Query expansion: We also used the crawled titles to perform query expansion. Again, based on the initial Lucene ranking, we used the URL titles found within the top K tweets (K = 10, 30, or 50) to extend the queries. Note that the top tweets as ranked by Lucene display the highest similarity with the original query. Therefore, if K is too high, more terms that are non-relevant will

appear in the expanded queries. To clarify the advantages of query expansion in the context of microblog search, here is an example. The original query in topic 87 is "*chicken recipes*". For this query, the collected title terms of the top 30 tweets are "chicken" (10x), "recipes" (14x), "recipe" (2x), "better" (4x), "healthy" (2x), and "easy" (2x). These terms intuitively form a sensible query (like, e.g., "**[Easy | better] healthy** chicken **[recipes | recipe]**"). In the expanded queries, the terms were weighted according to the term frequencies in the titles. It is therefore possible that high quality queries will describe the particular topic. We intend, using this form of query expansion, to capture more hidden terms (named entities, nouns, verbs, news sources, etc.). However, based on the original Lucene ranking, the expanded query is not always 'better' than the original one. A counter example is the query "*Steve Jobs' health*" in topic 106. For this query, user wants to know about the health situation of Steve Jobs'. However, the returned title terms of top 30 tweets for this query are "steve" (7x), "jobs" (7x), "is" (2x), "of" (3x), "health" (1x), "apple" (5x), "destroyer" (2x), "creator" (3x), "composed" (2x), and "products" (2x). These terms can be formed another query like: "Steve Jobs is **[destroyer** and **creator]** of **apple** composed products". The new query can be interpreted differently with respect to the original query. Moreover, some topics like "*Chipotle raid*" (topic 80), "*Superbowl commercials*" (topic 99), etc. did not allow collecting any additional title terms. By expanding the original document and query with the collected title terms, we get the idea that it is indeed possible to discover more relevant tweets. However, the method now suffers from some obvious flaws (e.g., biased by the original Lucene ranking). Therefore, further research is required to explore the full potential of the approach.

By expanding the original document and query with the collected title terms, we get the idea that it is indeed possible to discover more relevant tweets. However, the method now suffers from some obvious flaws (e.g., biased by the original Lucene ranking). Therefore, further research is required to explore the full potential of the approach.

In future research, we also intend to make use of other external sources like Wikipedia, large newspaper website, etc. to add semantics to the original tweets and queries (e.g., see [11]). The effectiveness of the expansion might also be improved by means of techniques like relevance-based language models [6], temporally-biased expansion models [7], etc.

2.4 Logistic Regression

In this section, we would like to focus on the classification step, whereby the relevance of a tweet for a query is determined by applying logistic regression on the given topics for the 2012 Track, trained on the 2011 topics. As mentioned in the introduction, tweets often contain abbreviated or skeleton terms to describe feelings or ideas. For example, some users write, e.g., ^-^, or =^^= to express happiness.. In general, many of these terms are barely decipherable, let alone in official English. However, because messages with these low-quality terms may still contain valuable information, we need to extract as many features as possible, in order to improve the effectiveness of the search system.

We chose logistic regression as a classifier to estimate the (binary) relevance because it is a simple yet powerful technique, and apart from a predicted label, we can use the estimated probability of relevance, directly resulting from the algorithm, to propose a sensible ordering of the results, as in a simple learning-to-rank approach. The features could however be combined in many different approaches like inference networks [8], neural network [9], or other learning-to-rank approaches, see for instance [10].

In the following, we briefly describe the features that were extracted from tweets and appeared to be useful in discriminating between relevant and non-relevant tweets:

1) **text_score (d,q):** This is a Lucene score which is calculated between the text in tweet *d* and query *q*.
2) **textTitle_score (d,q):** The Lucene score between text in tweet *d*, if possible extended with the title tokens, and query *q*.
3) **textTitle_Qtop10_score (d,q):** The Lucene score between text in tweet *d* with the title terms, and query *q* is extended by using the URLs titles found within top 10 tweets.
4) **textTitle_Qtop30_score (d,q):** The Lucene score between text in tweet *d* with the title terms, and query *q* is extended by using the URLs titles found within top 30 tweets.
5) **textTitle_Qtop50_score (d,q):** The Lucene score between text in tweet *d* with the title terms, and query *q* is extended by using the URLs titles found within top 50 tweets.
6) **url (d) - binary:** Does tweet *d* contain any links?
7) **hashtag (d) - binary:** Does tweet *d* contain any hashtag?
8) **retweet (d) - binary:** Does tweet *d* repost any tweet in Twitter system?
9) **text_length (d):** Number of words appearing in tweet *d* without the title terms.
10) **avetext_length (d):** The average word length (i.e., number of characters) in tweet *d* without the title term expansion.
11) **textTitle_length (d**): Number of words in the expanded tweet *d*, i.e.,including the added title terms.
12) **avetextTitle_length (d):** The average word length in the expanded tweet *d*, i.e., including added title terms.
13) **query_length (q):** Number of terms of query *q* without the title tokens.
14) **avequery_length (q):** The average word length in query *q* without the title tokens.

Because of the limited time, we did not investigate the relationship between Twitter users. For example, if a twitter user has many 'followers', he/she is more likely to provide relevant tweets than other users who has very few followers. Our future work will take into account these aspects as well. Other potentially interesting features might be formed by clustering the textual data, e.g., to deal with very short tweets (which have very

low Lucene ranking, because they do not contain any query terms, or are ranked artificially high, if they do).

3. EXPERIMENTAL RESULTS

We briefly describe the Twitter data collection process, the training data set, our submitted runs, and our results as we received them from the Microblog Track organizers.

3.1 Twitter data

For this year, the organization used the Tweets2011 corpus, already released for last year's track. The data collection contained a list of tweet ID's for approximately 16 million tweets, sampled between January 23rd and February 8th, 2011. The Track also provided the *twitter-corpus-tools*, to help all the participants download the tweets directly from Twitter. All participating groups had to retrieve the data by themselves, directly from the Twitter API, which means that each participant has a more or less different tweet collection. We had to fetch the data in HTML format (as we have no access to the API that provides the more complete JSON tweet meta-data).

The data was crawled on a personal desk top (Intel i3 processor, 4GB of RAM) in May 2012. The received HTTP response codes are given below:

- Status *200* – a good tweet for downloading.
- Status *302* – a re-tweet will be downloaded via redirect.
- Status *403* – invalid request: Twitter refuses to respond.
- Status *404* – the requested resource could not be found.

Table 2 shows some details about our tweet collection data. The table includes the English tweets which were detected by means of the LangDetect toolkit.

Table 2: Tweets collection data summary (May 2012)

types	no. elements
tweets status *200*	13,762,808
tweets status *302*	744,461
tweets status *403* and *404*	1,621,972
tweets found	14,004,761
tweets null	2,124,480
English tweets	4,597,488
Total tweets corpus	16,129,241

3.2 Relevance judgment data

We used last year's topics and relevance judgments, which were released by TREC, for developing and training our system. The used relevant scores for the judged tweets are the following:

- *2* – a highly relevant tweet.
- *1* – a relevant tweet.
- *0* – a non-relevant tweet.
- *-2* – a spam tweet.

However, some relevant tweets had in the mean time disappeared, (tweets deleted by the Twitter users, or user accounts no longer exist). Table 3 provides details on the annotated data in the total collection. Table 4 summarizes the collected titles.

Table 3: Tweets training data set

tweet types	no. elements
highly relevant (2)	502
relevant (1)	2150
non-relevant (0)	44423
spam (-2)	47
Total tweets training data	47122

Table 4: Titles training data

no.tweets	ave.*URLs*	ave. titles collected
24467	0.99	0.69

3.3 Submitted runs

We submitted the following four official runs with different feature combinations.

1. *IBCN1* – This run does not include any external resource (i.e., URL page titles). Only basic features which are combined using logistic regression. The features used are 1, 6, 7, 8, 9, 10, 13 and 14 from Section 2.4.
2. *IBCN2* – Logistic regression on all our features (1-14) in the Twitter collections data, including scores for extended tweets and queries with collected URLs page titles.
3. *IBCN3* – We only combined the best features (as estimated on the 2011 data set) in order to improve the efficiency of our search system. This was done in the following heuristic manner: we successively added one feature at a time (in order of decreasing performance for the corresponding single feature system), tested the performance on last year's topics, only retaining the added feature if performance improved (i.e., if it didn't we discarded it before continuing with the next).
4. *IBCN4* – The run is the same with *IBCN3*, however, we ordered the top 2% tweets (that's 6.12 tweets per query on average for the 2011 relevance judgment data; 91.62 per query for the 2012 Microblog TREC) based on their recency.

For training and testing our system, we applied 10-fold cross validation in which the relevance judgment data was divided into 10 parts (9 used as training and 1 as test data, and swapping the test "fold" for each of the 10 possible folds). Results were then averaged over all 10 executions. Table 5 shows the P@5, 10, 15 and 30 of relevance judgment data for each run. As we expected, *IBCN3* outperforms the runs *IBCN1* and *IBCN2*. However, results of *IBCN4* are better than *IBCN3* in P@5, 10, 15 and worse than *IBCN3* in P@30. It proves that the the top relevant tweets are indeed more recent than the other tweets in the collections, although the relationship between *recency* and relevance remains a future topic of research.

Table 5: Result on 2011 relevance judgment data

	P@5	P@10	P@15	P@30
IBCN1	0.4000	0.3898	0.3578	0.2959
IBCN2	0.4367	0.4041	0.3904	0.3323
IBCN3	0.4612	0.4241	0.3931	0.3558
IBCN4	0.4979	0.4388	0.3932	0.3163

3.4 Results

Table 6 gives details of the P@30, R-precision and MAP for our system in response to the 2012 topics, for each run. *IBCN2* outperforms the other runs, especially *IBCN1* and *IBCN4*. This suggests that the document and query expansion, are indeed helpful in order to improve the effectiveness of the tweet ranking system. Since the *recency* of the tweets is not explicitly taken into account for the shown metrics, the evaluation results from *IBCN4* do not match the purpose of the submitted run. Note that we only used the title field from the referred web pages, in order to extend the tweet and the original query. However, the meta descriptor tag of the web pages, as well as other elements, could have been employed as well, and is kept for future work. Also a more specific analysis as to what extent either the document expansion or the query expansion yield more to the increased effectiveness, is left for future work.

The Microblog Track 2012 received 121 runs from 33 participating groups. From the overall results [15], no participant seems to have obtained significant improvements. A major reason may be the vocabulary mismatch between original query and tweets collection, which is difficult to overcome. From these results, we suggest that document expansion needs to be included from 'external evidence' like Wikipedia, News, etc. Moreover, using expanded queries does not always give better (additional) information than the original unexpanded query. These approaches are potential ways to develop the system.

Table 6: Result from Microblog TREC 2012

	P@30	R_prec	MAP
IBCN1	0.1469	0.1585	0.1096
IBCN2	0.1904	0.1727	0.1408
IBCN3	0.1825	0.1712	0.1399
IBCN4	0.1379	0.1590	0.1190

4. CONCLUSION AND FUTURE WORK

In this paper, we described our approaches to developing a search system for the Microblog Track 2012. Based on the existing system, we intend to test novel and effective approaches to improve our results in the future.

We used a logistic regression classifier based on features extracted from the tweets, to predict their relevance. These tweet features describe local characteristics of the tweets, but also include ranking coefficients for the tweets, in combination with data crawled from the links provided in the tweets. In the future, we plan to investigate the influence of specific new features, and to use more advanced retrieval methods and unsupervised machine learning methods in order to capture more relevant terms in queries and topics.

Acknowledgements

This research was supported by Ghent University and iMinds (the former IBBT) in Belgium. We would like thank our colleagues who kindly helped us in crawling the Twitter corpus.

5. REFERENCES

[1] https://sites.google.com/site/microblogtrack/2012-guidelines
[2] http://twitter.com
[3] http://lucene.apache.org/
[4] C. J. Peng, K. L. Lee, G. M. Ingersoll. An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research 2002.
[5] http://code.google.com/p/language-detection/
[6] V. Lavrenko and W. B. Croft. Relevance-based language models. In Proc. 24th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 120–127, 2011.
[7] K. Massoudi, M.Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In Proc. 33rd Eur. Conf. on Information Retrieval, ECIR'11, pages 362-367, Berlin, Heidelberg, 2011. Springer-Verlag.
[8] H.Turtle and W. B. Croft. Evaluation of an interence network-based retrieval model. ACM Trans. Information Sys., 9(3):187-222, 1991.
[9] H. Takaki. Introduction to Fuzzy Systems, Neural Networks, and Genetic Algorithms. Intelligent System. Ch. 1, pp. 1-33, 1997.
[10] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. Hullender. Learning to Rank using Gradient Descent. In Proc. 22nd Int. Conf. Machine Learning. Bonn, Germany, 2005.
[11] E. Meij, W. Weerkamp, M. de Rijke. Adding Semantics to Microblog Posts. Proc. WSDM 2012.
[12] http://shuyo.wordpress.com/
[13] H. Kwak, C. Lee, H. Park and S. Moon. What is Twitter, a social network or a news media. Proc. WWW. 2011, pages 591-600, North Carolina, USA. ACM.
[14] R. Nagmoti, A. Teredesai, M. De Cock. Ranking Approaches for Microblog Search, Proc. ACM WI-IAT'10.
[15] I. Soboroff, I. Ounis, C. Macdonald, J. Lin. Overview of the TREC-2012 Microblog Track. In Proc. TREC 2011.

Table 7: Examples of crawled page titles in Twitter data

URLs	Title link
http://bit.ly/euuqdD	Makeup Man on February 11
http://bit.ly/euUqt7	ECOtality integrates Blink Network with Cisco's solution
http://bit.ly/eUUu1A	Expatriate Tax Senior Director job BDO USA LLP New York NY Indeed.com
http://bit.ly/euUUjM	Asian Cup 2011 Song Qatar 2011 YouTube
http://bit.ly/euzwQt	State of Green Business 2011 MNN Mother Nature Network
http://bit.ly/euZXzG	bestarticlepublisherlisting.info The Leading Best Article Publisher Listing Site on the Net